Transparency and Reliability in the Data Supply Chain

Groth, P.T.

published in
IEEE Internet Computing
2013

DOI (link to publisher)
10.1109/MIC.2013.41

document version
Early version, also known as pre-print

Final version appears in IEEE Internet Computing
Groth, Paul, "Transparency and Reliability in the Data Supply Chain," Internet Computing, IEEE, vol.17, no.2, pp.69,71, March-April 2013 doi: 10.1109/MIC.2013.41

Transparency and Reliability in the Data Supply Chain

Paul Groth
Network Institute, VU University Amsterdam

Linked data is enabling the creation of large-scale distributed data supply chains. However, it is often unclear the origins of the information these supply chains produce. Unlike coffee, we don't have a fair trade certificate for data. Here, I outline how standards and research in data provenance are leading towards of "fair trade" data.

Keywords: provenance, data, supply chain, linked data, transparency

On 26 November 2012, several major technology blogs, including Techcrunch (http://techcrunch.com) and Gizmodo (http://gizmodo.com), reported that Google was taking over ICOA, a public Wi-Fi hotspot provider, for US$400 million, causing ICOA's stock price to surge[1]. However, these reports were based on a false press release that somehow made its way to PRWeb, a press release distributor. Later, the mistake was caught and the blogs updated their posts. This example illustrates how errors within the supply chain of information can easily propagate and have a dramatic impact in the real world. What these blogs failed to do was adequately check where the information had come from **in the first place.**

Indeed, we often judge quality by relying on knowing something's origins and how it was produced, termed *provenance*. For example, we can tell coffee's quality by knowing where it came from, the roasting process it underwent, and how it was brewed. Knowing provenance also lets us give credit to the actors in a system: whether it's the barista, the roaster, or the bean farmer. This same approach applies to information.

On the Web, we publish lightweight forms of provenance intended for humans all the time. On Twitter, for example, marking provenance arose organically. Authors started using shorthand such as "RT" to denote a retweet or "MT" to denote that a tweet was modified, as well as the "@" symbol to refer to the original author (see Figure 1). On blogs, using blockquotes is common to demarcate that certain information comes from another source, and we use hyperlinks to attribute that information.

Figure 1. Provenance within a tweet. This tweet illustrates a number of items used to denote provenance. One sees the original author of the tweet (ISWC2013), that the tweet was retweeted (RT @iricelino) by the person identified by @iricelino and that it was subsequently retweeted by Paul Groth.

Although these conventions are useful for people to determine provenance and thus judge information's

[1] Hesseldahl, Erik "Google Sources Say Company Didn't Buy ICOA Wireless" http://allthingsd.com/20121126/google-sources-say-company-didnt-buy-icoa-wireless/. Accessed January 7, 2013.

quality, they're less useful when we move to the data arena. Quality and credit concerns only increase in this environment, where data is processed not only by humans but also by machines.

The Data Supply Chain

Just as transnational shipping has made it possible to spread the production of goods across long, distributed supply chains involving various suppliers and services, the Internet and *linked data* standards are enabling a distributed information supply chain. Essentially, RDF and HTTP provide the standard containers for shipping and interchanging data. However, most *big data* doesn't come fully formed in one monolithic bloc but is instead smaller pieces combined. Twitter isn't one stream but billions of individual messages taken together. Linked data itself is, per definition, not a cohesive uniform dataset but an amalgamation of multiple data sources ranging from biology to movies. In addition, many of these sources have been translated and transformed from their original version to make them easier to integrate.

A good example of a data supply chain is in pharmacology: data ranges from excel spreadsheets produced from experimental lab work, to data from specialized commercial curation services that read the literature and extract structured data, to information organized by public organizations. These data are then intermixed using data warehousing pipelines that normalize data to common schemas or are combined using ad hoc analysis procedures.

Such pipelines are common across domains and raise questions about the quality, reliability, and transparency of the results that they produce. Several common questions arise:

- Who is responsible for an error in the analysis' results?
- Do we have a license to the information, and what does this license permit?
- Is this information derived from curated sources or an automated process?
- Where are the bottlenecks in our pipeline?
- Why is our information out-of-date?
- Is this information from an authority?

These questions are similar to those that were asked about the incorrect blog posts, but answering them is harder because of the complexity, automation, and scale of data supply chains.

Provenance for a Transparent Data Supply Chain

These questions are beginning to be answered through advances in tracking and using data provenance. One work enumerates a large set of requirements for provenance systems.[1] Here, I highlight some current approaches to tackling these requirements and outline where the challenges still lie.

For Web information, work on information diffusion and cascades[2] is exposing how text and ideas are reused and repurposed across social networks. These approaches use textual similarity and other clues to determine how phrases propagate and mutate through social media. Indeed, work on information cascades is increasingly important in studying social structures and Internet marketing. The New York Times Research & Development Lab has built its own system, called Cascade (http://nytlabs.com/projects/cascade.html), using information cascades to see how its work affects the larger social media conversation.

In the data arena, the community is attacking the problem from different ends. We've become quite good at tracking provenance in single systems that have been specifically instrumented to acquire it — for instance, by modifying the operating system, redesigning the database, instrumenting computational workflow systems, or tracking computations in frameworks such as Hadoop. These systems are good evidence that we can process big data transparently.

However, data supply chains are inherently distributed systems that extend across application and organizational boundaries. To address this challenge, the community is developing standards to let systems transfer and interchange provenance information. The W3C has recognized this need and set up a working group to create a recommendation for how to interchange provenance between systems. A key target for this specification is data made available on the Web. The working group is already at the last stages of its

process and has developed a common provenance interchange specification, PROV,[3] that covers common provenance components such as attribution, processing activities, and derivation. A key part of this specification's design is to support a spectrum of use cases from simple information about who has responsibility for data to complex descriptions of how data has been manipulated and combined. Moreover, a set of implementations is emerging that support this specification (www.w3.org/2011/prov/wiki/ProvImplementations). Just as the barcode helps track products through the physical supply chain, PROV allows us to track information products across the data supply chain.

Challenges

Although the availability of a standard, software, and methodologies for provenance capture and exposure is encouraging, challenges remain, particularly with respect to the data supply chain.

Provenance Reconstruction

Although standards are a crucial component to tracking the data supply chain, we still need mechanisms to capture this information when the software and processes involved haven't been specifically instrumented. Just as auditors often reconstruct what's happened in a traditional supply chain, we need systems to do this for the data supply chain. Thus, a key area of research is reconstructing data's provenance. This work can build on notions from information diffusion, but must extend to cover complex data processing as well as data's heterogeneous nature.

Fair Trade Data

Given the data supply chain's length and the complexity of the procedures involved, the provenance of any one result can be huge and easily overwhelm users. A key challenge is to develop coherent abstractions for provenance that provide insight into the data's quality on the basis of how it was produced. Additionally, we need good mechanisms for communicating the resulting summaries. Essentially, what we need is a fair trade certificate for data — a seal of approval that says our data is produced and derived in a way that we as data consumers think is correct.

The Web and specifically linked data have enabled the production of complex data supply chains in which the component parts are packaged, processed, and cleaned by distributed sets of outsourced providers. These supply chains are often too opaque and too complex for humans to find the errors. Advances in data provenance tracking and standardized interchange are making it possible to create more transparent supply chains. Although challenges remain, I'm hopeful that in the coming years, our data will have provenance as good as that of our coffee.

References

1. P. Groth et al., "Requirements for Provenance on the Web," *Int'l J. Digital Curation*, vol. 7, no. 1, 2012, pp. 39–56.
2. D. Gruhl et al., "Information Diffusion through Blogspace," *Proc. 13th Int'l Conf. World Wide Web*, ACM, 2004, pp. 491–501.
3. P. Groth and Luc Moreau, "PROV-OVERVIEW: An Overview of the PROV Family of Documents," W3C working draft, 11 Dec. 2012; www.w3.org/TR/2012/WD-prov-overview-20121211/.

Paul Groth *is an assistant professor at the VU University Amsterdam and is affiliated with its Network Institute. His research interests include* ***data provenance, web science, and knowledge integration*** *Groth has a* ***Ph.D*** *in computer science from the University of Southampton Contact him at p.t.groth@vu.nl*